Evaluation of different biological data and computational classification methods for use in protein interaction prediction

Supplementary Material

1. Feature Attributes

1.1. List of Attribute Groups

Table S1. A total of 162 features are characterized into 17 categories. Two styles of feature encoding are used to result in very different sizes of the feature vectors. The second column lists the numbers of features when using "Detail" encoding for each category. In the "summary" encoding, this number would be 1. (Table 3 in the paper presents the coverage of each attribute group used in our feature set.)

GROUP INDEX	# OF FEATURES	DATASET	Attribute Property	REFERENCE	NOTE
1	20	Gene Expression	Real value: [-1, 1]	[8]	Co-Expressed Score
2	21	GO Molecular Function	{1, 0}	[7, 15]	Co-Function Score
3	33	GO Biological Process	{1, 0}	[7, 15]	Co-Process Score
4	23	GO Component	{1, 0}	[7, 15]	Co-Location Score
5	1	Protein Expression	Real Value – Non Negative	[9]	Co-Expressed Score
6	1	Essentiality	{2 , 1, 0}	[14]	
7	1	HMS_PCI Mass	{ 1, 0}	[5, 3]	Matrix model for co-complex and co-pathway prediction. Spoke model for direct PPI prediction [3]
8	1	TAP Mass	{ 1, 0}	[4, 3]	
9	1	Y2H	{ 1, 0}	[1, 2, 3]	
10	1	Synthetic Lethal	{ 1, 0}	[10, 13]	
11	1	Gene Neighborhood / Gene Fusion / Gene Co-occur	{ 1, 0}	[10]	
12	1	Sequence Similarity	Real value - Non negative	[15]	
13	4	Homology based PPI	Discrete: Non-negative (Most 0, 1)	[15, 16]	
14	1	Domain-Domain Interaction	Real value between [0, 1]	[11]	Co-Domain Score
15	16	Protein-DNA TF group binding	Non-negative discrete, most 0	[6]	Co-Binding Score
16	25	MIPS Protein Class	{ 1, 0}	[17]	
17	11	MIPS Mutant Phenotype	{ 1, 0}	[17]	

1.2. Details about Each Attribute

1.2.1. Gene Expression Data

The gene expression data were obtained from ref. [8] and contained 20 gene expression datasets recorded under more than 500 conditions (each measuring a time series expression profile) was downloaded from http://www.psrg.lcs.mit.edu/Networks/data/expressionData.txt). We can either compute one global similarity score (under "Summary" encoding) for each pair of proteins or 20 distinct scores (under "Detailed" encoding) for each pair.

- In summary encoding, we calculated, for each pair, the Pearson correlation value considering all conditions and used it as one attribute.
- In detail encoding, we split the 500+ set into the following subsets: 20 subsets based on their experimental sources and conditions based on the criteria given in http://www.psrg.lcs.mit.edu/Networks/data/expressionData.txt. We then calculated the Pearson CC for each dataset and therefore obtained 20 features for this group.

1.2.2. SGD's Gene Ontology (Co-function, Co-process, and Co-localization)

Gene Ontology (GO) based information was downloaded from SGD [7] and include:

- molecular function of a gene product,
- biological process in which the gene product participates,
- cellular component where the gene product.

- In summary encoding, for each pair in each of the three GO hierarchies trees, we use as feature the value of how many times both are in the same category. This results in three values as attributes. We treat the functional catalog as a hierarchical tree of functional classes. Each protein is either a member or not a member of each functional class, such that each protein describes a "subtree" of the overall hierarchical tree of classes. The "functional similarity" between two proteins is defined as the frequency at which the intersection tree of the two proteins occurs in the distribution. Intuitively, the intersection tree represents the function shared by the two proteins. Finally, a single real value is derived to represent this similarity for a protein pair.
- In detail encoding, we generate each GO-feature as a discrete feature {0 or 1}: "1" means, both proteins share the same function /component /process. "0" means otherwise. There are 34 types of processes, 22 types of function, 24 and types of component features. Each class was mapped to one binary variable ("attribute").

1.2.3. Protein expression data

Ghaemmaghami, et al [13] presented experimental protein abundance data for yeast. Since this data set includes just one condition's expression, we used the absolute difference as our protein co-expression attribute.

- In summary encoding: Due to there is only one condition expression in this data, we use the absolute difference of the protein expression value.

- In detail encoding: Due to there is only one condition expression in this data, we also use the absolute difference of the protein expression value. So here the detailed encoding is the same as the summary encoding for this feature.

1.2.4. Essentiality

1106 ORFs are listed in the essential ORF list, downloaded from www.sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt. Based on the advice by the authors of this feature set, we assume that anything not listed can be considered to be nonessential (NE). Any gene deemed essential (E) is one that cannot be made into a haploid or homozygous deletion strain. The co-essential feature is a 3-value categorized feature: 0 means NE/EN, 1 means NN, 2 means EE)

- In summary encoding: This is a one value feature.
- In detail encoding: This is a one value feature. Here the detailed encoding is the same as the summary encoding for this feature.

1.2.5. High throughput direct PPI data set

Two types of high throughput direct data were used, (1) derived from mass spectrometry and (2) from Y2H screens:

- Mass spectrometry data: These experiments use individual proteins as 'hooks' to biochemically purify protein complexes. The identity of the proteins located in these complexes is then determined by mass spectrometry. TAP [4] (tandem affinity purification) and HMS-PCI [5] (high-throughout mass-spectrometry protein complex identification) are two of the protocols used for this technique. Both protocols may miss true complexes when the affinity is weak or transient or when the tagged protein may be misfolded or its interaction capability disturbed by the tag. We used TAP and HMS-PCI as separate attributes. To convert complex relationships to interaction pairs, we use the spoke model [3] for the direct protein-protein interaction prediction task, resulting in 3224 pairs for TAP (spoke) and 3618 pairs for HMS-PCI. For the other two tasks, we employed the matrix model to use these two mass spectrometry features.
- Y2H (yeast two-hybrids screen) data: In the Yeast two-hybrid system, potential pairs of proteins are expressed as two separate fusion (hybrids) proteins in yeast that are brought together by the DNA-mediated interaction of the fusion proteins. Therefore, this method requires that the two test proteins are capable of interacting in the environment of the nucleus. Thus, some proteins which are natively localized in other compartments of the cell may fail to interact. 5614 Y2H interactions were downloaded from [3].

- In summary encoding: For each highthroughput experiment, the values are determined by the experiments. We do not have calculation processing here.
- In detail encoding: The values are determined by the experiments. Here the detailed encoding is the same as the summary encoding for this feature.

1.2.6. Synthetic Lethal

The synthetic lethal data described as {0, 1} discrete feature pairs were derived from the union of the following data sets:

- 295 synthetic lethal interaction from the first high-throughput study on genetic interactions in yeast [13a]
- 591 synthetic lethal interactions parsed from MIPS were downloaded from http://mips.gsf.de/proj/yeast/tabels/interaction/genetic_interact.html.
- A genetic interaction network containing approximately 1000 genes and approximately 4000 interactions [13]:

- In summary encoding: The values are determined by the experiments. We do not have calculation processing here.
- In detail encoding: The values are determined by the experiments. Here the detailed encoding is the same as the summary encoding for this feature.

1.2.7. Sequence Derived Features

This attribute is the union of the following three data sets described by Mering, et al [10]:

- Conserved gene neighborhood: 42 sequenced genomes were searched for instances of conserved neighborhood between genes.
- Co-occurrence of genes: Each entry in the orthology-database COG9, derived from 42 completely sequenced genomes. We used the Mering et al.-derived feature of potential pairs of genes with mutual information higher than 0.5 (close matches to the 13 most frequent patterns were ignored, as they are mostly phylogenetic) [10].
- Gene fusion events: These were detected by the presence of a gene in more than one COG cluster. Single fusion events were not considered significant.

- In summary encoding: We download the data from [10]. Due to each method's low coverage, we use the union of them as the feature.
- In detail encoding: Here the detailed encoding is the same as the summary encoding for this feature.

1.2.8. Sequence Similarity

This feature was obtained from the SGD NCBI-BLASTP: ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence_similarity/ [15]. We only used the yeast to yeast alignment result from this database. All BLASTP hits obtained with the default parameters that had E-value less than or equal to 0.01 were used and the query protein was excluded from the results.

- In summary encoding : This is a one value feature.
- In detail encoding: This is a one value feature. Here the detailed encoding is the same as the summary encoding for this feature.

1.2.9. Homology Based PPI

As for the feature "sequence similarity", we also used the SGD NCBI PSI-BLAST hits results [15] to derive the homology feature. We use the 0.001 as a cutoff on the E-value to decide the homology pairs. In this case, every ORF in S. cerevisiae was queried using PSI-BLAST against NCBI's non-redundant (nr) protein database subset of four species:

- Caenorhabditis elegans (WormBase) ➔ 19844 hits
- Drosophila melanogaster (FlyBase) ➔ 26268 hits
- Homo sapiens (ENSEMBL (HUMAN) ➔ 101066 hits
- Saccharomyces cerevisiae (SGD) ➔ 30489 hits

The final features were obtained by determining if a candidate Yeast protein-protein pair interacts in other species or not. If yes, the feature was the number of times their homology proteins found to interact, otherwise "0".

- In detail encoding: To each species above, we first search for the homology protein within that specie for a specific Yeast protein. Then for a candidate Yeast protein pair, if we could find the homology proteins for both proteins in another species. We check to see if these two homology proteins interact or not (For worm, fly and human, we use DIP [6] to check. For Yeast, we use Y2H to check).
- In summary encoding: We use the union of the above four features in the summary encoding style.

1.2.10. Domain-domain interaction feature

Deng et al [11] used maximum likelihood estimation methods to infer interacting domains based on sequence analysis. They use yeast two-hybrid protein interaction data and treat protein sequences as "bag of domains". The data were downloaded from http://www.cmb.usc.edu/msms/ProteinInteraction/ using the file ProteinAction_025_80SGDY2H.txt (trained based on the Y2H PPI). We used the protein interacting probability derived from the above derived domain-domain interaction probability as features.

- In summary encoding: This is a one value feature.
- In detail encoding: This is a one value feature. Here the detailed encoding is the same as the summary encoding for this feature.

1.2.11. MIT Gene Regulator Binding Data

Transcription Factor (TF) binding data were downloaded from http://jura.wi.mit.edu/young_public/regulatory_code/GWLD.html [6]

- In summary encoding, we used a p-value cutoff to define binding. For each pair of proteins, we have counted the number of transcription factors that bind to both genes, and have used this number as the attribute.

- In detail encoding, we group the TFs based on the MIPS protein class catalog into 16 TF groups. For each TF group, we counted the number TFs that bind to both genes, and used this number as one of our attributes. This resulted in 16 features in this group.

1.2.12. MIPS Derived Features

We employed the two functional properties in MIPS, (1) protein class catalog and (2) protein knock out mutant phenotype catalogs:

- MIPS protein class catalog: Yeast proteins are assigned to different protein classes according to the MIPS Protein Class Catalogue. For each protein class, we recorded pairs of proteins that both fall into that class. The data were downloaded from ftp://ftpmips.gsf.de/yeast/catalogues/protein_classes/
- MIPS knock out phenotype: Mutant phenotypes for yeast genes were obtained from the MIPS Mutant Phenotype Catalogue. For each mutant phenotype, we recorded pairs of proteins whose encoding genes both have that mutant phenotype. The data were downloaded from ftp://ftpmips.gsf.de/yeast/catalogues/phenotype/

- In summary encoding, features correspond to how many times a pair in each of the two MIPS property trees belong to the same category. This results in 2 values from two catalogs.
- In detail encoding, we define each MIPS second level property class as a {0 or 1} discrete feature. "1" means the two proteins belong to the same class or mutant phenotype. "0" means otherwise. There are 25 protein classes and 11 first level mutant knock out phenotype classes.

2. Precision vs. Recall Curves of Six Different Classifiers

We used two measures to evaluate the performance of a classifier: (a). Precision vs. Recall Curves and (b). R50 Partial Area under Receiver Operator Characteristic Curves.

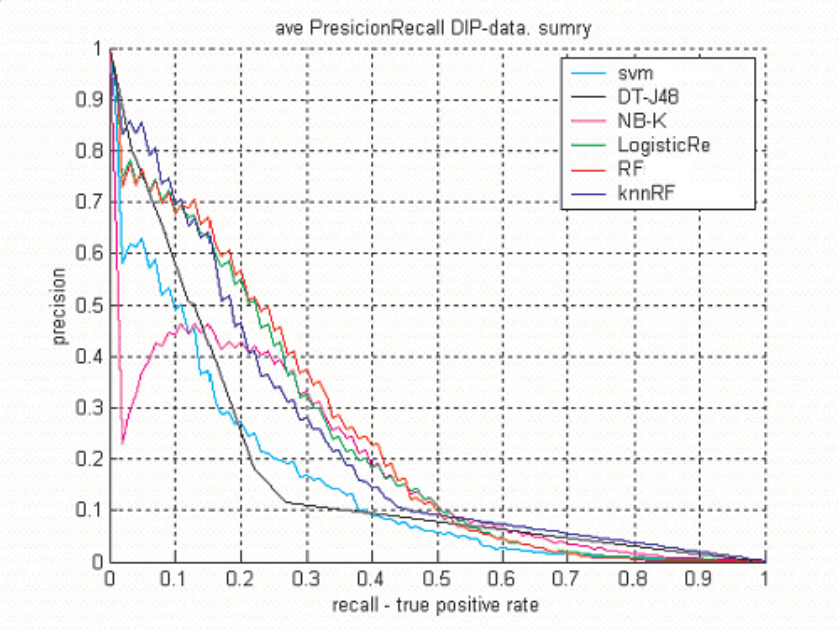
- In Precision vs. Recall curves, precision refers to the fraction of interacting pairs predicted by the classifier that are truly interacting ("true positives"). Recall measures how many of the known pairs of interacting proteins have been identified by the learning model. The precision vs. recall curve is then plotted for different cutoffs on the predicted score.
- Receiver Operator Characteristic (ROC) curves plot the true positive rate against the false positive rate for the different possible cut-off values of the predicted score. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. It can take values from 0.0 to 1.0. In our prediction task, we are predominantly concerned with the detection performance of our models under conditions where the false positive rate is low. Here, we use 50 as a cut-off, i.e. R50 is a partial AUC score that measures the area under the ROC curve until reaching 50 negative predictions. Here we also show the R25, R100 and R150 values.

Considering the highly skewed distribution of interacting and non-interacting pairs, we employed a cost-sensitive strategy. This strategy assumes that the classifiers pay higher costs if a positive example is misclassified into the negative class. Based on this assumption the classifier then moves the prediction boundary to minimize the training costs. Then, we have a cost-sensitive parameter to choose during the modeling training. Like other parameters, this parameter is also chosen by various train-test experiments to find the best value to use.

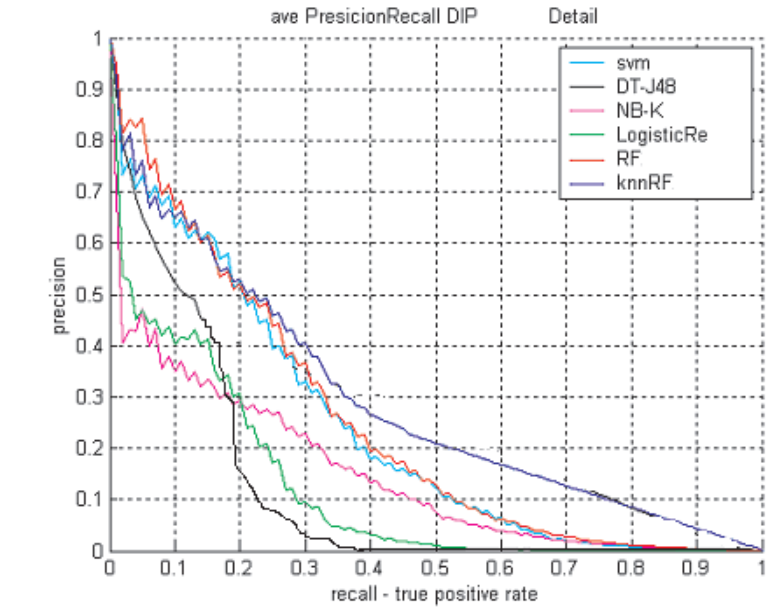
Below is the supporting material for Figure 3 of the paper. Provided are the precision vs. recall curves of the physical interaction task and the co-pathway tasks by the six classifiers and two feature encoding types.

2.1 Physical interaction task (DIP)

2.1.1. Summary Encoding

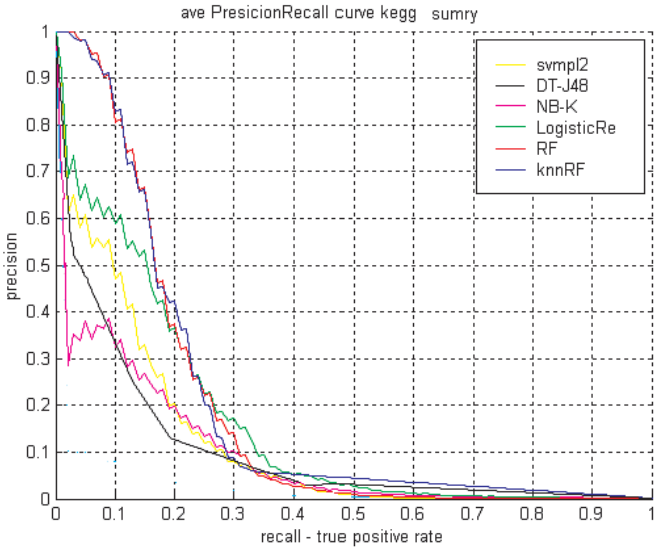


2.1.2. Detailed Encoding

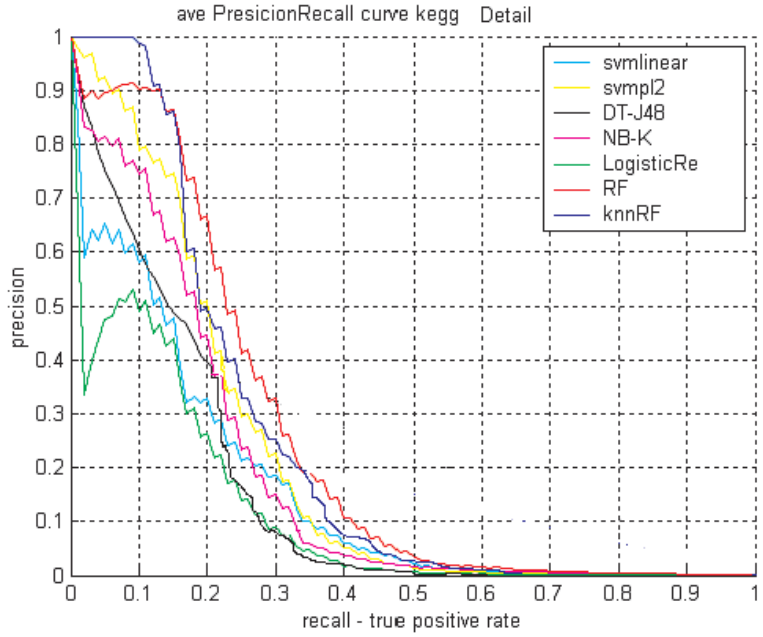


2.2 Co-pathway membership task (KEGG)

2.2.1. Summary Encoding



2.2.2. Detailed Encoding



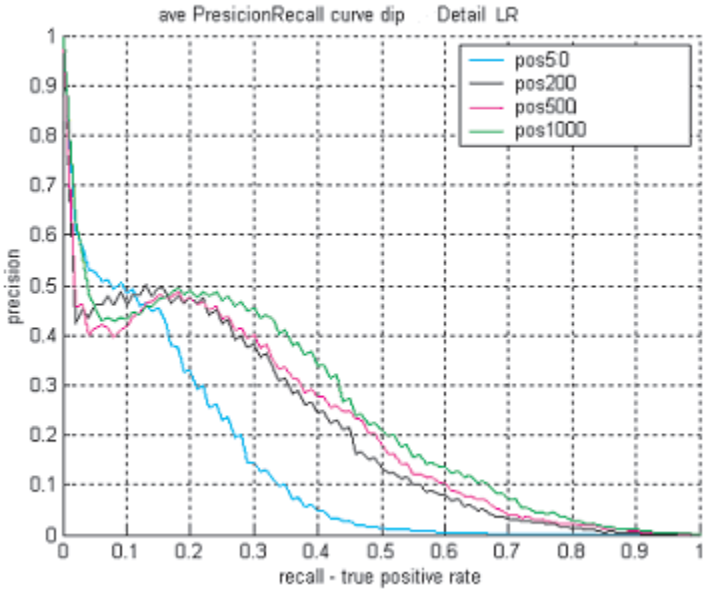
3. Classifier Performance Depending on Training Data Size

3.1 Logistic regression Precision vs. Recall curves when changing training size

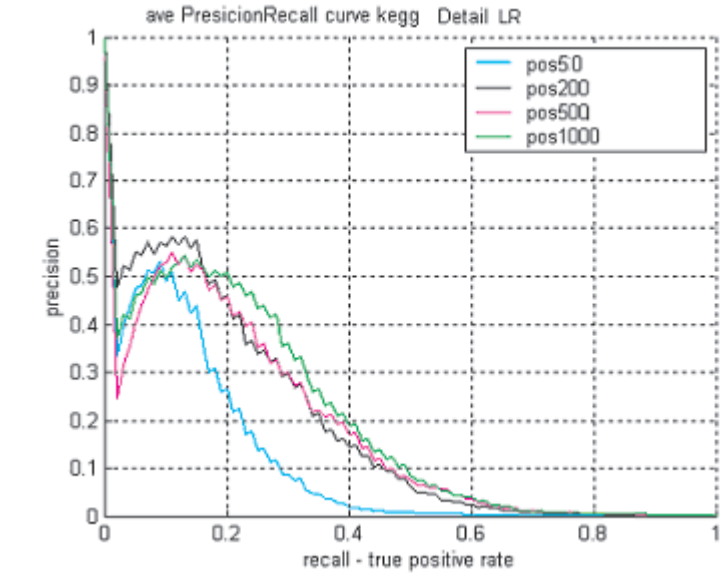
Below is the supporting material for Figure 4 of the paper. Provided are the precision vs. recall curves using Logistic Regression (LR) applied to the physical interaction task and the co-pathway tasks when changing the size of the training set. Supporting the conclusion from Figure 3 of the paper, these curves show that the LR performance is worse than the RF performance, even when increasing the train size drastically.

Precision vs. Recall curves were obtained by varying the training set size to include (a) ~50 interaction pairs. (b) ~200 interaction pairs. (c) ~500 interaction pairs. (d) ~1000 interaction pairs. Features were encoded as "Detailed".

3.1.1 Physical interaction task (DIP)



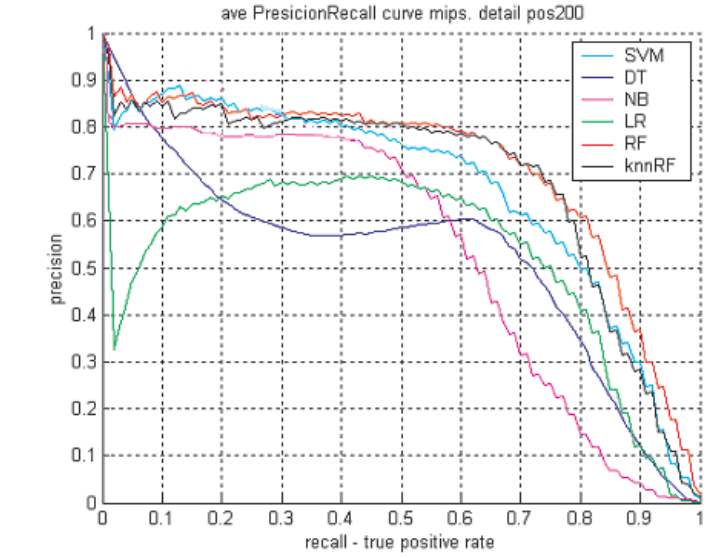
3.1.2 Co-pathway membership task (KEGG)



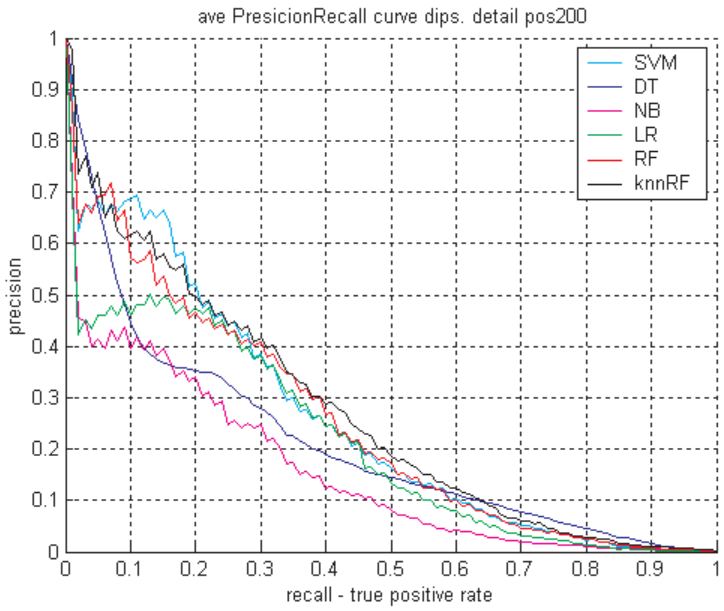
3.2 Six classifiers Precision vs. Recall curves when changing training size

Furthermore, we also make the comparison between all six classifiers when the training set contains around 120,000 examples. The precision vs. recall curves of all six classifiers upon increasing the size of the training set to ~120,000 with ~200 interacting pairs are shown next. Features were encoded as "Detailed". From the Precision-Recall curves, it can be seen that even when we use a larger training set, the RF classifier is still ranked among the top two best methods in all tasks.

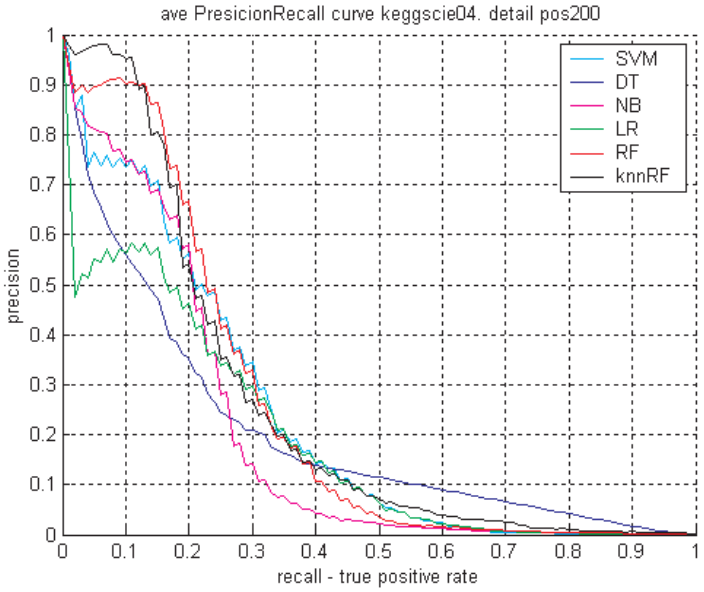
3.2.1 Co-complex prediction task (MIPS)



3.2.2 Physical interaction task (DIP)



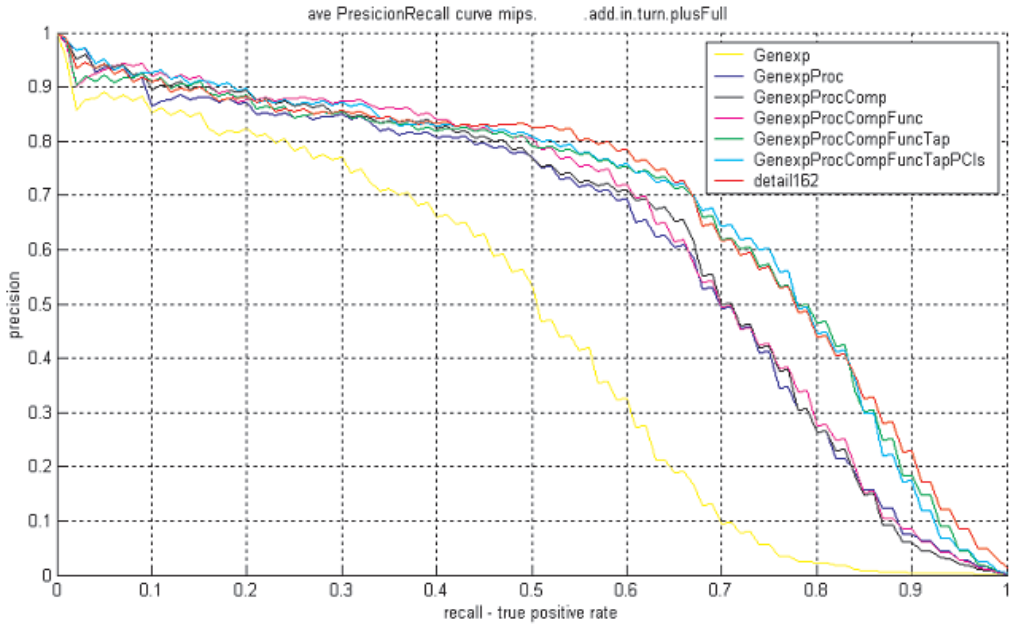
3.2.3 Co-pathway membership task (KEGG)



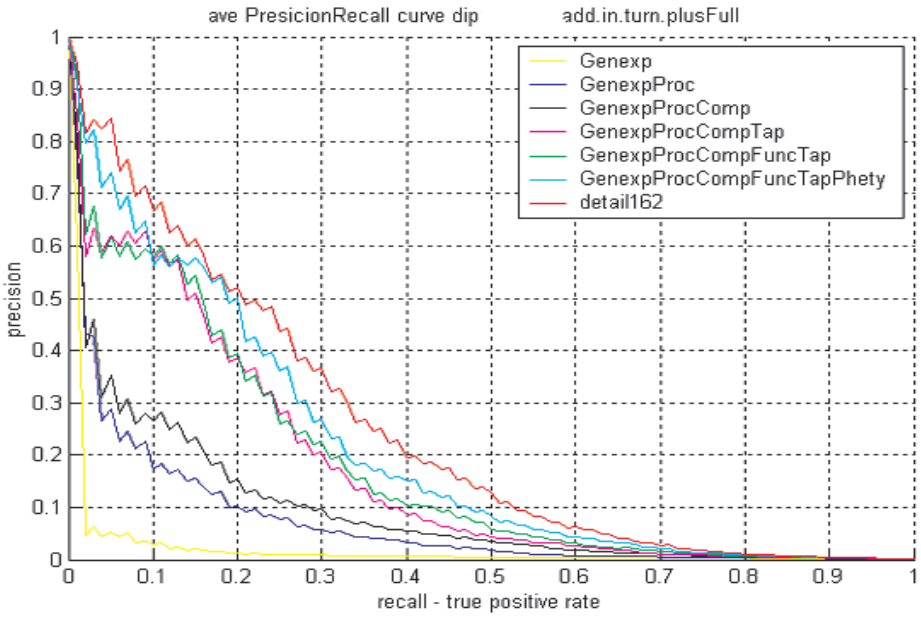
4. Performance Considering Feature Composition

Shown below is the performance comparison when using the top 6 ranked feature categories for each prediction task. The features were added one after the other according to the order given in Table 4 of the paper. The RF classifier with "Detailed" feature encoding was used for this experiment. Each curve represents the score using all features up to that rank (1 to 6). The seventh curve presents the Precision-Recall curve when using the full set of features.

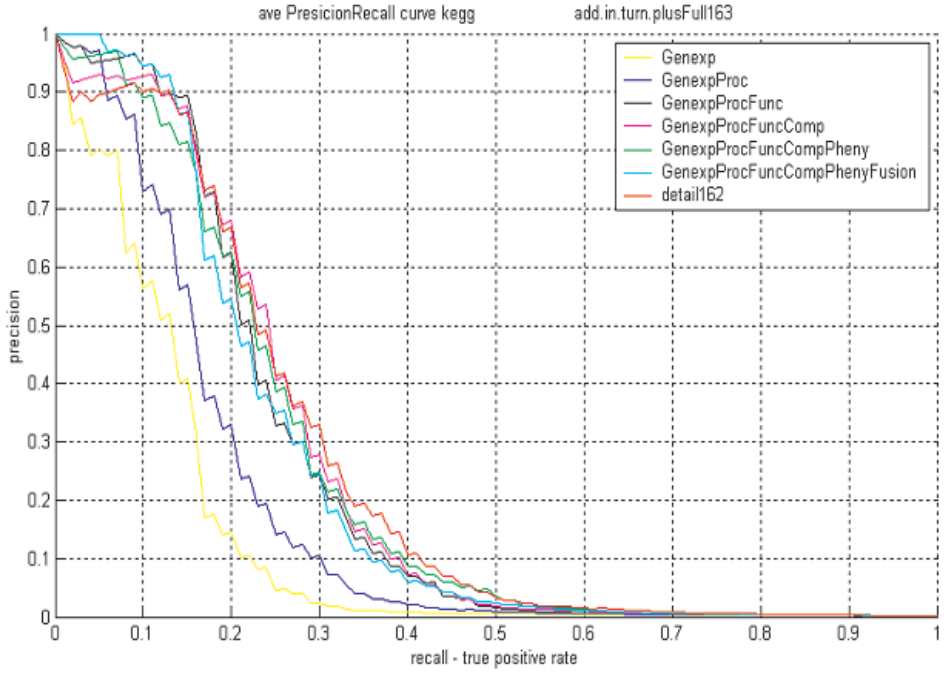
4.1. Co-Complex Task (MIPS)



4.2. Direct Protein-Protein Interaction (DIP)



4.3. Co-pathway Task (KEGG)



5. References:

1. Uetz P, et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 403(6770):623-7. (2000)

2. Ito T, et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome., Proc Natl Acad Sci U S A. 10;98(8):4569-74. (2001)
3. Bader GD, Hogue CWV. Analyzing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology 20:991-997 (2003)
4. Gavin AC, et. al, Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 415(6868):141-7. (2002)
5. Ho Y, et al., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180-3. (2002)
6. Harbison,CT., Gordon,DB., Lee,TI., Rinaldi,NJ., Macisaac,KD., Danford,TW., et. al, (2004) Transcriptional regulatory code of a eukaryotic genome. Nature, 431, (7004):99-104.
7. The Gene Ontology Consortium, (2000) Gene Ontology: tool for the unification of biology. Nature Genet.25 25-29 (2004 Nov version).
8. Z. Bar-Joseph*, G. Gerber*, T. Lee*, N. Rinaldi, J. Yoo, F. Robert, B. Gordon, E. Fraenkel, T. Jaakkola, R. Young, and D. Gifford Computational discovery of gene modules and regulatory networks. Nature Biotechnology, 21(11) pp. 1337-42, 2003
9. Ghaemmaghami S, et al., Global analysis of protein expression in yeast. Nature. 425(6959):737-41. (2003)
10. von Mering C, et al., Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399-403. (2002).
11. Deng M, et al., Inferring domain-domain interactions from protein-protein interactions. Genome Res. 12(10):1540-8. (2002)
12. Dolinski,K., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Saccharomyces Genome Database (SGD). http://www.yeastgenome.org (2004)
13. (a) Tong A.H.Y. et al, Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants, Science 294, 2364-2368 (2001)
13. (b) Tong A.H.Y. et al. Global Mapping of the Yeast Genetic Interaction Network. Science. 303: 808-813. (2004)
14. The Saccharomyces Genome Deletion Project: http://www-sequence.stanford.edu/group/yeast_deletion/project (2004 Nov version)
15. Dolinski,K., Balakrishnan,R., Christie,K.R., Costanzo,M.C., et. al, (2004) "Saccharomyces Genome Database" (2004 Nov version).
16. Xenarios I, et al., DIP: The Database of Interacting Proteins: 2001 update, Nucleic Acids Res. 29(1):239-41 (2001)
17. Mewes,HW., Amid,C., Arnold,R., Frishman,D., Guldener,U., et. al}, MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res.32, (Database issue):D41-4 (2004)
18. Huh,WK., Falvo,JV., Gerke,LC., Carroll,AS., Howson,RW., Weissman,JS., O'Shea,EK., Global analysis of protein localization in budding yeast. Nature, 425, (6959):686-91. (2003)